W308–W313 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 15 May 2014*
doi: 10.1093/nar/gku369

AIDA: *ab initio* domain assembly server

Dong Xu[1], Lukasz Jaroszewski[1,2], Zhanwen Li[1] and Adam Godzik[1,2,3,*]

[1]Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA, [2]Center for Research in Biological Systems, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0446, USA and [3]Center of Excellence in Genomic Medicine Research (CEGMR), King Fahad Medical Research Center, King Abdulaziz University, P.O. Box 80216, Jeddah 21589, Kingdom of Saudi Arabia

Received March 3, 2014; Revised April 14, 2014; Accepted April 15, 2014

ABSTRACT

AIDA: *ab initio* domain assembly server, available at http://ffas.burnham.org/AIDA/ is a tool that can identify domains in multi-domain proteins and then predict their 3D structures and relative spatial arrangements. The server is free and open to all users, and there is an option for a user to provide an e-mail to get the link to result page. Domains are evolutionary conserved and often functionally independent units in proteins. Most proteins, especially eukaryotic ones, consist of multiple domains while at the same time, most experimentally determined protein structures contain only one or two domains. As a result, often structures of individual domains in multi-domain proteins can be accurately predicted, but the mutual arrangement of different domains remains unknown. To address this issue we have developed AIDA program, which combines steps of identifying individual domains, predicting (separately) their structures and assembling them into multiple domain complexes using an *ab initio* folding potential to describe domain–domain interactions. AIDA server not only supports the assembly of a large number of continuous domains, but also allows the assembly of domains inserted into other domains. Users can also provide distance restraints to guide the AIDA energy minimization.

INTRODUCTION

Structures and functions of protein domains are highly conserved. This applies to domains found in single-domain proteins or as parts of multi-domain proteins. Evolution of proteins, especially eukaryotic ones, can be viewed as combinatorial process where single-domain proteins are fused to perform functions that require two (or more) domains (1–3) and as the result, most eukaryotic proteins consist of two or more domains. However, due to technological limitations, solving structures of multi-domain proteins is difficult and only 1/3 of solved structures in the Protein Data Bank (PDB) (4) contain more than one domain. Typically, either a single-domain (usually prokaryotic) version of the protein or a single domain excised from a large, multi-domain protein is targeted for structure determination. This leads to a situation where we often have experimentally solved or accurately predicted structures of individual domains in a multi-domain protein, but do not have the structure of the entire protein chain. For multi-domain proteins, template-based protein prediction algorithms typically only output models of individual domains as full-length templates are simply not available. Those that do provide full-length models, usually assemble them in the completely random arrangement.

While formally similar to the protein–protein docking problem, domain assembly has a much smaller search space due to the chain connectivity constraints between consecutive domains. At the same time, interactions between domains can be described by the same energy terms as interactions within folded domains. Therefore, we could use protein folding potential to guide domain assembly. Our benchmark tests have shown that *ab initio* knowledge-based folding potential (5) used in the *ab initio* domain assembly (AIDA) program not only can guide the simulation to find the correct domain positions, but also can help the selection of the best assembled multi-domain protein model with high success rate (6).

AIDA server can predict the assembly of proteins with any number of continuous domains and proteins containing domains inserted into other domains (discontinuous domains). AIDA server also provides an option of modeling and assembly of multi-domain proteins starting directly from the provided sequences. In that case, a protein is iteratively split into domains by aligning them with the modeling templates found by the FFAS-3D (7) fold recognition program. Furthermore, AIDA server supports restraint-guided domain assembly following optional user-specified inter-domain distance restraints.

*To whom correspondence should be addressed. Tel: +1 858 646 3168; Fax: +1 858 795 5249; Email: adam@godziklab.org

AIDA INPUT

The default input for the AIDA server is the primary sequence of the entire protein chain and 3D coordinates of individual domains in the PDB format. Residue numbering in the domain structures should be consistent with that in the sequence. Otherwise the numbering in the final model of the whole structure could contain errors. For similar reasons the server also does not accept gaps in the domain models. Therefore, we recommend that the users renumber and model the missing fragments of individual domains' structures before submitting them to the server. Both steps can be performed semi-automatically by modeling programs or online servers such as Modeller (8) or ProtMod (http://ffas.burnham.org/protmod-cgi/protModHome.pl), which would not only build the missing domain fragments but may also renumber the residues (they usually accept starting residue number for the sequence alignment). At the same time we want to note that user-provided domain structures do not have to cover the entire sequence. The sequence fragments that correspond to regions between provided domain structures are interpreted as linkers and are subject to movements. It also means that residues in the terminal regions of the domain structures could be truncated to elongate the flexible linker regions.

The situation where all structures of domains are known is not typical. In fact, for most proteins domain boundaries and structures of individual domains are unknown. In such situations, the AIDA server can accept a protein sequence as the only input. The server will then perform domain splitting, modeling and assembly automatically. AIDA domain assembly program has no limit of the maximum sequence length. However, the server relies on PSIPREDs (8) secondary structure prediction result, which accepts sequences with up to 10 000 residues.

Another option of the server allows a user to upload distance restraints, i.e. the distances between pairs of Cα atoms from different domains. This approach is useful in situations when global structure of a multi-domain protein is difficult to solve, but some key interactions between domains can be determined experimentally (e.g. by nuclear magnetic resonance spectroscopy) or can be predicted [for instance by the analysis of correlated mutations (9) or using structural fragments (10)]. Such information about inter-domain contacts would help AIDA simulation program to find the best domain arrangement, which fulfills the distance restraints.

AIDA PROCEDURE

The complete procedure of AIDA domain assembly is illustrated in Figure 1. In the first step, AIDA server performs PSI-BLAST (11) search against non-redundant protein database clustered at 85% to generate the log-odds profile (position-specific substitution matrix). This profile is then used by PSIPRED program to predict protein's secondary structure. Predicted secondary structure and PSI-BLAST output are then used to predict solvent accessibility for all residues using a two-layer neural network.

Linker regions (or 'linkers') comprise segments between adjacent domains. They usually have no regular secondary structures and their lengths vary in different proteins. By

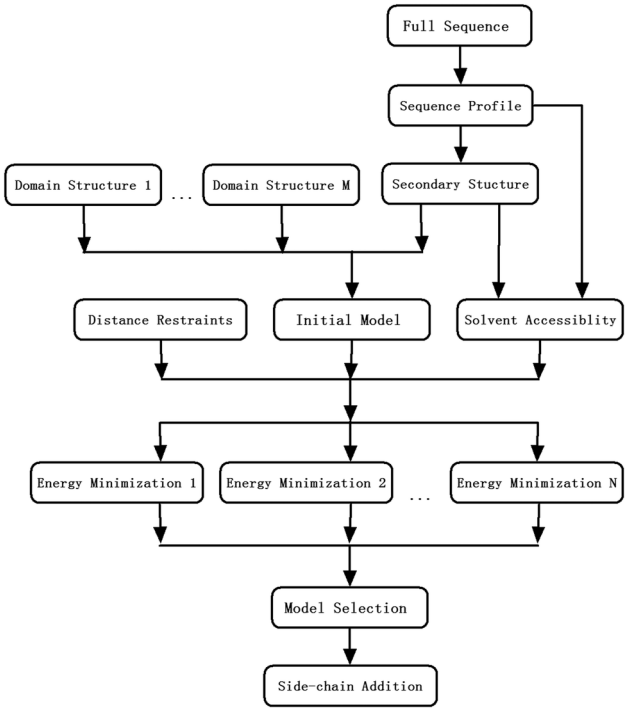

Figure 1. Flowchart of domain assembly procedure implemented in AIDA.

default, AIDA program uses four residues around each domain boundary as the linker region (i.e. two residues at each end of each domain). However, some linkers are much longer and, even if they are included as termini in the domain structures, their conformations are not always accurately determined. In such cases, users may remove terminal residues before uploading domain structure files to AIDA and they will be modeled as the linker region. Longer linker will result in the increase of the search space of the AIDA optimization.

In the next step, AIDA algorithm calculates bond lengths, bond angles and torsion angles for all the residues in the domain structures. Coordinates of the missing residues (i.e. linker regions) are generated based on the secondary structure types predicted by PSIPRED. From those geometrical features, the initial full-length model is built in the torsion-angle space, which ensures that the domains are connected.

In the subsequent optimization step, AIDA uses knowledge-based statistical potentials, which have been tested for *ab initio* protein folding (5). In this model, protein structure is represented only by main-chain atoms and side-chain centers. Hydrophobic interaction described as a difference of predicted and actual solvent accessibility, is one of the most discriminative terms in domain–domain assembly. The residues which are predicted to be buried (have low predicted solvent accessibility values) but are exposed to the solvent in the current domain arrangement, are drawn towards other domains by the AIDA optimization protocol. If a user has uploaded the distance restraints, difference of pairwise distance in the current assembled model and that provided by the user will be used as one additional energy term to guide the simulation.

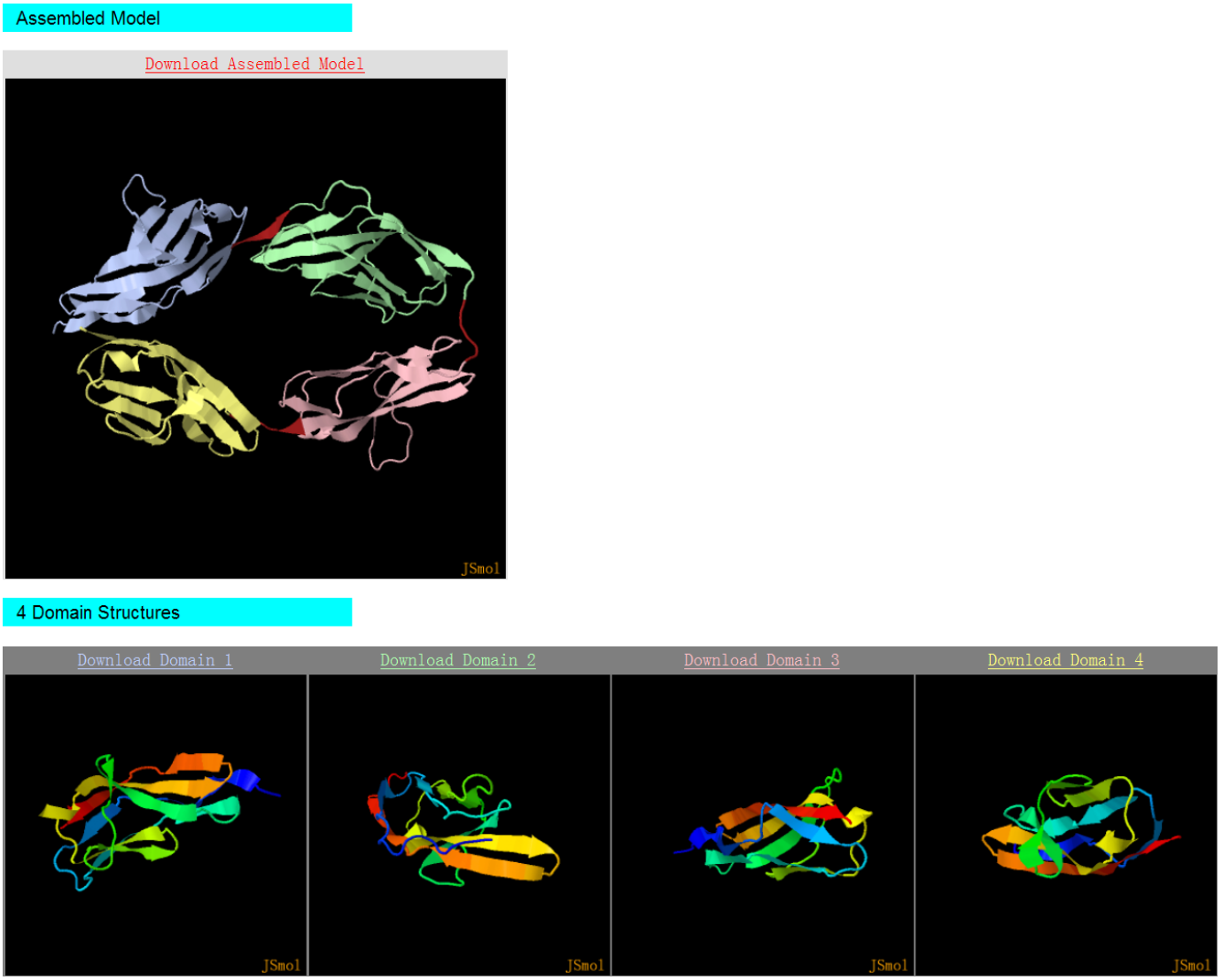

Figure 2. Assembly result for four continuous domains of N-terminal fragment of axonin-1 from chicken.

From the initial model, AIDA attempts to perturb the linker conformation, which results in the different arrangement of domain structures. Linker conformation change (i.e. movement) is accepted only if the total energy of the new full-length model is lower than that of the previous model. The program will stop if the total number of attempts exceeds $1000L$ (where L is the length of the entire sequence) or 200 consecutive movements have failed (indicating that the model reaches the global/local minimum state).

We generate 50 different full-length models, from trajectories that start from different initial random numbers. The assembly with the lowest total energy is selected and then SCWRL4 (12) program is used to add side-chain atoms to generate all-atom model as the final output.

USE CASES

Assembly of multiple continuous domains

If a user uploads several domain structures corresponding to continuous regions in the uploaded protein, then AIDA performs flexible assembly of continuous domains. The server allows up to 20 continuous domains.

An example of the result of this procedure performed for four domains of N-terminal fragment of axonin-1 from chicken (PDB: 1cs6, chain A) is shown in the upper panel of Figure 2. There are almost no interactions between the first and second domains and between the third and fourth domains. The second and third domains have several interactions. The first and fourth domains are predicted to be close to each other despite the fact that they are separated by the long distance in sequence. This domain arrangement generated by AIDA is roughly the same as in the experimental full-length structure [TM-score (13) of the assembled model

Figure 3. Assembly result for one discontinuous domain containing inserted domain of phosphoglycerate kinase from *Trypanosoma brucei* bisubstrate analog.

versus experimental structures is 0.64]. For comparison, individual domains are shown in the bottom part of the figure.

Assembly of discontinuous domains with inserted domains

Large insertions and deletions are frequently seen in proteins' evolution. In such events, short motifs or even an entire domain could be inserted into loops in existing domains. As a result, domains that are compact in 3D structures, become discontinuous at the sequence level. Standard flexible assembly model is not applicable here since it would not retain relative orientation and position of the two parts of the discontinuous domain (it would interpret them as two different domains).

In fact, each of the two segments separated by the inserted domain could contain part of a domain or even multiple domains. However, in AIDA protocol they are described as the first and third domains. Similarly, the middle inserted part, even if it contains multiple domains, is treated as a single (second) domain. When we build the initial full-length model, only the second domain is built in the torsion-angle system, which keeps it connected with the first domain at one end but may be distant from the third domain at the other end. Then, during the AIDA simulation, the energy term that penalizes the broken chains gradually pulls the second and third domains together.

An example of the assembly of discontinuous domain with inserted domain in phosphoglycerate kinase from *Trypanosoma brucei* bisubstrate analog (PDB: 16pk, chain A) is shown in Figure 3. The assembled model is shown at the upper part of the figure while separate domains are shown at the lower part of the figure. (The first and third regions are shown together since they form one structural domain.) Since there are two short linkers connecting the two domains, sampling space for domain–domain interaction is limited. The assembled model is very close to the native structure with TM-score = 0.75.

Automated prediction of multi-domain protein structures

AIDA server can also perform fully automated prediction of structures of individual domains and assemble the structure of a multi-domain protein starting directly from the protein sequence. In this case, FFAS-3D fold recognition program is used for template detection. Since FFAS-3D uses local–local variant of dynamic programming algorithm, aligned regions include only domains with significant similarity to the query. As a result, the protein is automatically split into at most three parts (the regions aligned with the templates, and, possibly, unaligned N- or C-terminal region or both). The model of the aligned region is built using Modeller (14) while FFAS-3D is applied again to the unaligned regions. This procedure is continued iteratively until there is no remaining unaligned region longer than 20 amino acids or all unaligned regions do not contain any predicted regular secondary structures. Note that domain division here is based on the alignment with PDB templates. Even if part of the sequence is aligned with a multi-domain structure, we still treat this region as one domain. That is to say, domain arrangement for those continuous domains is fixed unless users build their models off-line and provide them as separate domains to the server (using domain assembly option).

We have tested the AIDA protocol on the set of targets in CASP10 (10th community wide experiment on the critical assessment of techniques for protein structure prediction). All the single-domain proteins were accurately identified as such, and domain boundaries were correctly predicted for over 85% (21 out of 24) multi-domain proteins. Some CASP10 targets contain domains with no modeling templates available in PDB, but even then their boundaries were usually correctly identified based on the alignment of the neighboring domains.

Figure 4 shows one example of automated multi-domain structure prediction for hypothetical protein BT_2966 (locus tag NP_811878.1) from *Bacteroides thetaiotaomicron* VPI-5482 genome (15). The primary sequence as well as the predicted secondary structure types and their confidence scores are shown on the top of the figure. After running FFAS-3D against the whole protein, C-terminal domain of putative chitobiase from *Bacteroides thetaiotaomicron*

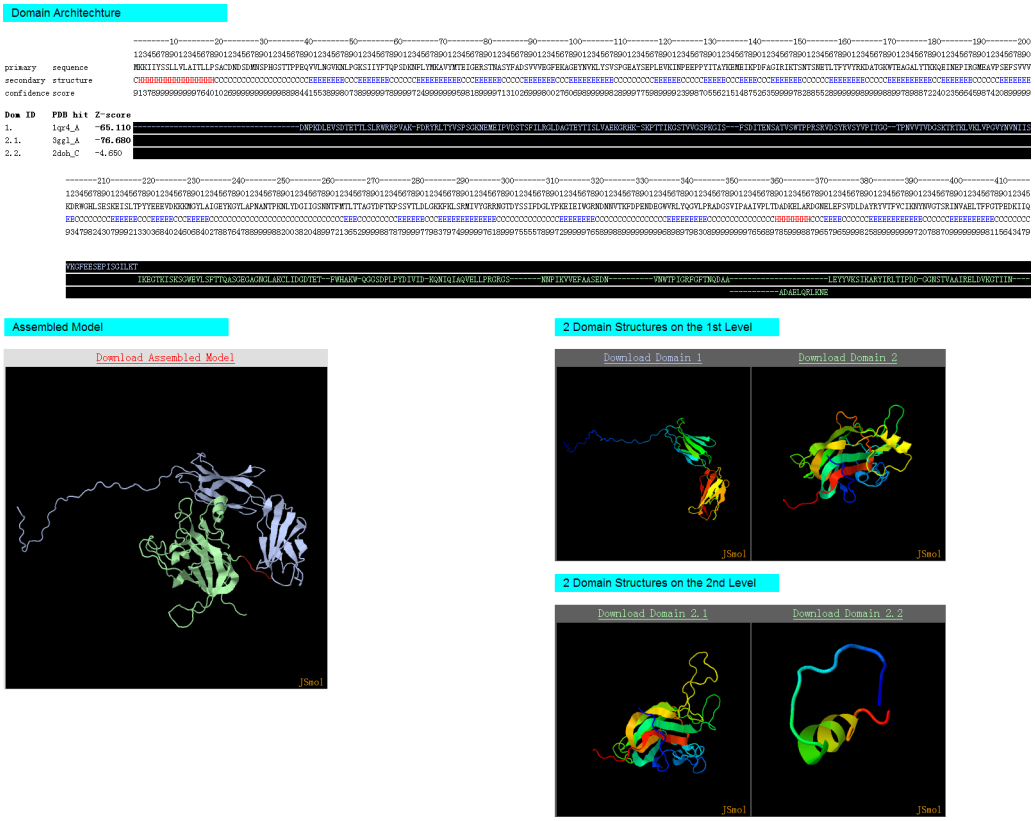

Figure 4. Example of automated domain splitting, modeling and assembly for hypothetical protein BT_2966 from *Bacteroides thetaiotaomicron* VPI-5482.

(PDB: 3ggl, chain A) was selected as the best template. The initial alignment included C-terminal part of the protein and, thus, effectively divided the sequence into two parts, with predicted domain boundary at 216. FFAS-3D score for the alignment was below the cutoff value of −34.0, indicating significant similarity. For the unaligned region in the N-terminal region, FFAS-3D was performed again and two fibronectin type-III domain segment from chicken tenascin (PDB: 1qr4, chain A) was selected as the best template. 37 residues in the N-terminal still have no alignment. However, predicted secondary structure indicates that this part is only composed of a long helix and a long coil, suggesting that it is probably a signal peptide. Hence, no further template identification was performed for this region. Modeller was then used to model full N-terminal region encompassing residues 1–216 (the N-terminal 37-residue fragment was built as a long coil).

There is a short insertion in the middle of the second domain, which contains 22 amino acids and is predicted to contain a helix. We ran FFAS-3D for this region, which identified angiogenesis inhibitor, angiostatin (PDB: 2doh, chain C) with high Z-score. In the bottom-right of Figure 4, the Modeller model for the whole C-terminal is labeled as Domain 2.1 while the Modeller model for the inserted short motif is labeled as Domain 2.2. The coil built by Modeller for this inserted part of Domain 2.1 was then removed and discontinuous domain assembly was performed for them. The assembly result is the final Domain 2. At last, continuous domain assembly was run for the two domains and the final full-length model was generated.

The iterative protocol is completely automated. Sometimes, one large domain may contain multiple insertions. In such cases, AIDA will perform the discontinuous domain assembly for them one by one.

CONCLUSIONS

The AIDA server performs assembly of multiple domain structures, including inserted domains. Possible domain–domain interactions could be derived from the assembled structure. It also supports the automated multi-domain protein structure prediction, i.e. it helps users to predict domain boundaries, build 3D models for individual domains and finally assemble them together.

It has to be noted that in cases when two or more consecutive domains do not have detectable templates, then domain boundaries based on FFAS-3D alignments would be most likely incorrect. In such cases, domain boundaries can be predicted using external resources, which contain domain definitions independent from 3D structures [for example the Pfam (16) database].

ACCESSION NUMBERS

PDB: 1cs6, chain A, 16pk, chain A, 3ggl, chain A, 1qr4, chain A and 2doh, chain C.

FUNDING

National Institute of Health [GM087218]. Funding for open access charge: National Institute of Health [GM087218] and institutional funds.
Conflict of interest statement. None declared.

REFERENCES

1. Vogel,C., Bashton,M., Kerrison,N.D., Chothia,C. and Teichmann,S.A. (2004) Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.*, **14**, 208–216.
2. Zmasek,C.M. and Godzik,A. (2012) This déjà vu feeling–analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Comput. Biol.*, **8**, e1002701.
3. Gerstein,M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.*, **3**, 497–512.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
5. Xu,D. and Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, **80**, 1715–1735.
6. Xu,D., Jaroszewski,L., Li,Z. and Godzik,A. (2014) AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. *submitted for publication*.
7. Xu,D., Jaroszewski,L., Li,Z. and Godzik,A. (2014) FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics*, **30**, 660–667.
8. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
9. Marks,D.S., Colwell,L.J., Sheridan,R., Hopf,T.A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.
10. Xu,D. and Zhang,Y. (2013) Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*, **81**, 229–239.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Krivov,G.G., Shapovalov,M.V. and Dunbrack,R.L. Jr. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.
13. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
14. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
15. Xu,J., Bjursell,M.K., Himrod,J., Deng,S., Carmichael,L.K., Chiang,H.C., Hooper,L.V. and Gordon,J.I. (2003) A genomic view of the human-bacteroides thetaiotaomicron symbiosis. *Science*, **299**, 2074–2076.
16. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* . (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.